Jurnal Mantik
Volume 4, Number 3, Nov 2020, pp. 1573-1581 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

# Employee Turnover Analysis Using Comparison of Decision Tree and Naive Bayes Prediction Algorithms on K-Means Clustering Algorithms at PT. AT

Silfanus Kingki Setianto, Dwiki Jatikusumo

[1]Program Study Teknik Informatika,Fakultas Ilmu Komputer, Universitas Mercu Buana, Jl. Raya Meruya Selatan,Kembangan, Jakarta, 11650, Indonesia

E-mail: kingkii.setianto@gmail.com, dwiki.jatikusumo@mercubuana.ac.id

**ARTICLEINFO**

Article history:
Received: 12/07/2020
Revised: 22/08/2020
Accepted: 30/09/2020

**Keywords:** *Data mining, Prediction, Turnover, Naïve Bayes Algorithm, Decision Tree, Rapid Miner 5.1.*

**ABSTRACT**

PT. AT is a leading information technology service provider company in Indonesia. Being one of the leading companies in Indonesia, of course, is greatly supported by the quality of employees and the number of employees working there. One way to increase the great influence in a company's development is to retain qualified employees and continue to develop them. So that employee turnover is one thing that needs to be watched out for. Of course, we can predict this, one of which is by considering the number of active and non-active employees based on the average number of each year. This study study the level of resigned employees using data mining years from 2015 - 2019 obtained from Human Capital in the company. The data obtained will be clustered according to the division and level of employees and then analyzed using the Decision Tree and Naïve Bayes methods. So from the results of the analysis, it can be used as a reference to predict employee turnover rates in 2020, so that this research is expected to help in developing plans for employees in the company in order to achieve company targets. It is hoped that the results of this study will be used to anticipate employee turnover that is not controlled.



## 1. Introduction

An organization is formed to achieve common goals, but to achieve goals effectively, good and correct management is needed. There are various opinions about the notion of management, although it basically has the same meaning. Management really determines an organization in achieving an organizational success. Management includes the process of planning, organizing, directing, and using supervision of the efforts of other organizational members so that they can achieve predetermined organizational goals [1].

PT. AT is a leading information technology service provider company in Indonesia, which has its head office in Tangerang. This company was founded in 2002, which has provided integrated IT solutions to large companies in Indonesia. And has served businesses ranging from Banking to Telecommunications, Pharmacy and even in the field of Government with timely implementation results and sufficient budget. Starting from a company consisting of 20 dedicated people, now PT. AT has developed into one of the leading IT companies in Indonesia. Since its formation, PT. AT aims to provide high value and excellent service as a customer competitive advantage leaps.

To maintain the integrity and good values of the company, of course, cannot be separated from the hard work of employees. Then employee job satisfaction in an organization is very important role in order to create good performance. Employees who have high job satisfaction have better performance in carrying out their duties than those who are dissatisfied with their work. The factors that influence employee job satisfaction are employee turnover, motivation and job enrichment [1]. However, over time, employee turnover cannot be avoided. Therefore this research was conducted to analyze how high the percentage level of employee turnover. The observation method used is by taking the employee data available at Employee Services and then processing it using a rapid miner by adding a data processing algorithm. The first step when the data has been obtained is pre-processing, namely cleaning the data, so that the data is perfect and intact. Then from the clean data, preliminary processing is carried out, namely clustering using the k-means algorithm, then comparing predictions using the Naïve Bayes algorithm and the Decision Tree, to find the most accurate algorithm for processing the data. in order to get perfect and complete data. Then from the clean data, initial processing is carried out, namely clustering using the k-means algorithm, then comparing



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

predictions using the Naïve Bayes algorithm and the Decision Tree, to find the most accurate algorithm for processing the data. in order to get perfect and complete data. Then from the clean data, initial processing is carried out, namely clustering using the k-means algorithm, then comparing predictions using the Naïve Bayes algorithm and the Decision Tree, to find the most accurate algorithm for processing the data.

## 2. Literature Review

Data mining devices are activities that include the collection and use of historical data that finds regularities, patterns and relationships in large data sets. The purpose of this definition is the process of searching for previously unknown information from large data sets.

Decision-making in a fairly complex scope such as in civil engineering, really depends on the cognitive abilities of the decision maker, so that it can directly impact the quality of the decision itself. All data must be taken into consideration in making decisions using various generally accepted methods. This data will be useful information if it can be interpreted correctly. Data mining has solidified into a form of concentration in artificial intelligence science. Since 1960 Data Mining has been continuously developed to maximize the capabilities of databases whose capabilities and capacities also continue to increase rapidly. The development of Data Mining has been able to encourage the use of raw data in databases to become a very meaningful source in the formation of various forecast models. The database will remain only raw data if there is no interpretation of the data that has been collected. The ability of Data Mining to accurately model interpretation because it is supported by the capacity of Data Mining to read very large data [16]. With the Data Mining model, no matter how large the available database is, it will be able to be retrieved, tabulated, processed and then the interpretation model is carried out properly in accordance with generally accepted statistical principles. As well as live surveys, questionnaires, and visual tabulations, historical pavement performance will support Data Mining's ability to help develop a practical pavement management system. The Data Mining Model developed in this study is expected to be able to perform interpretations that can be used as initial data to carry out maintenance and improvement of optimal road quality.

K-Means Clustering is a data analysis method or Data Mining method that performs the modeling process without supervision (unsupervised) and is a method of grouping data using the partition system. K-means is a non-hierarchical clustering method that seeks to partition data into in a cluster / group so that data that has the same characteristics are grouped into the same cluster by first determining the number of clusters [5].

To process the K-means Clustering algorithm data, the data starts with the first group of randomly selected centroids, which are used as the starting point for each cluster, and then performs an iterative (iterative) calculation to optimize the centroid position.

The Naive Bayes algorithm is a classification method using probability and statistical methods proposed by the British scientist Thomas Bayes [18]. The Naive Bayes Algorithm predicts future opportunities based on past experiences so it is known as Bayes' Theorem. The main characteristic of this Naïve Bayes Classifier is a very strong (naïve) assumption of independence from each condition / event [3].

Naïve Bayes for each decision class, calculates the probability on the condition that the decision class is correct, given the object information vector [18]. This algorithm assumes that the attributes of an object are independent. The probability involved in producing the final estimate is calculated as the sum of the frequencies from the "master" decision table.

Decision tree is a flow-chart like tree structure, where each internal node shows a test on an attribute, each branch shows the result of the test, and leaf node shows the classes or class distribution [8].

The decision tree technique is generally used in research operations, especially in the analysis of decisions that require several strategies to achieve certain objectives. Another use of the decision tree is as a descriptive mean to calculate conditional probabilities. The definition of a decision tree is a decision support tool that uses a decision model and its possible consequences, including the outcome of the opportunity event, resource costs, and utility.

A decision tree consists of a structural tree with features on the leaves / nodes. It is usually considered unvaried because each node contains one feature. Branching starts with the feature that has the highest weight, descending to the lowest weight that is built recursively in dividing and conquering ways. One of the most useful characteristics of a decision tree is its comprehensiveness, which can be easily traced.

## 3. Research methods

In this study, an experiment was carried out using employee data owned by the company. The research stages include data collection, pre-processing, processing using research algorithms, and data analysis in



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

order to find the correlation between attributes to determine the most influential attributes in this study. Here are the starting stages of the research described above.

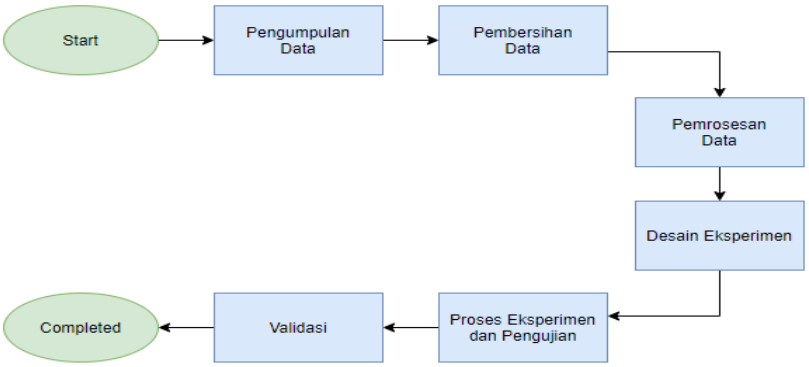


**Fig 1.** Research methods

### 3.1 Data collection

The method of collecting data is by requesting employee data that is owned by the company to Employee Services, the requested data is open to internal employees. The data collected is data on active and inactive employees from 2015 to 2019. The following is an example of data collected, before any processing is carried out.

| | Gender | Date Of Join | Date Of Resign | Employee Status | Durasi | >12 Bulan | Job Title |
|---|---|---|---|---|---|---|---|
| 2 | Female | 27-Nov-15 | 26-Feb-16 | Non Active | 2 | no | Corporate Finance ... |
| 3 | Female | 2-Feb-15 | 30-Apr-15 | Non Active | 2 | no | EXECUTIVE ASSIS... |
| 4 | Female | 31-Aug-15 | 31-Oct-15 | Non Active | 2 | no | EXECUTIVE ASSIS... |
| 5 | Male | 2-May-18 | 2-Aug-18 | Non Active | 3 | no | Section Head of E... |
| 6 | Female | 16-Nov-15 | 15-Mar-16 | Non Active | 3 | no | Internal Audit Officer |
| 7 | Male | 17-Jan-17 | 21-Apr-17 | Non Active | 3 | no | Dept. Head of Cosp... |
| 8 | Female | 1-Nov-17 | 28-Feb-18 | Non Active | 3 | no | EXECUTIVE ASSIS... |
| 9 | Female | 9-Nov-15 | 29-Feb-16 | Non Active | 3 | no | Internal Audit Officer |
| 10 | Male | 1-Aug-16 | 30-Nov-16 | Non Active | 3 | no | Internal Audit Officer |
| 11 | Female | 1-Aug-16 | 30-Nov-16 | Non Active | 3 | no | Accounting Supervi... |
| 12 | Female | 28-Sep-15 | 31-Dec-15 | Non Active | 3 | no | Business Develop... |
| 13 | Male | 3-Apr-17 | 11-Aug-17 | Non Active | 4 | no | Section Head of Int... |
| 14 | Female | 2-Feb-15 | 15-Jun-15 | Non Active | 4 | no | Section Head of Pa... |
| 15 | Female | 27-Jul-15 | 30-Nov-15 | Non Active | 4 | no | Section Head of Pe... |
| 16 | Female | 1-Nov-17 | 31-Mar-18 | Non Active | 4 | no | Section Head of Re... |
| 17 | Male | 1-Dec-16 | 5-May-17 | Non Active | 5 | no | Head of Finance |
| 18 | Female | 19-Sep-16 | 15-Mar-17 | Non Active | 5 | no | Section Head Quali... |

**Fig 2.** Data collection

The attributes or fields of the data include:

a) Gender : As the name implies, this field contains gender data.
b) Date Of Join : Contains data on the employee's entry date
c) Date of Resign : Contains exit date data (*resign*) the employee
d) Employee Status : Contains current employee status data, whether *active* or non-active
e) Duration : This field contains data on how long the employee isan already resigned person to work at the company
f) > 12 Months : This field is in addition to increasing accuracy in data processing, useful for assisting data filtering, so that it can produce more accurate accuracy. \
g) Job Title : This field contains the title data of the employee

### 3.2 Data Pre-Processing

At this stage, data cleaning is carried out, by filling in several empty fields (replace missing values) with the values that have been determined, so that the data is clean, so that when processed using a rapid miner, the results are more accurate. Then determine which field will be the label, or key point for processing. The following is a table list column and labeling for processing this data.

**Table 1**
Data Pre-Processing

| No. | Column Name | Type | Label |
|---|---|---|---|
| 1 | Gender | Binominal | Attribute |
| 2 | Date of Join | Polynominal | Attribute |
| 3 | Date of Resign | Polynominal | Attribute |



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

| No. | Column Name | Type | Label |
|---|---|---|---|
| 4 | Employee Status | Binominal | Label |
| 5 | Duration | Integer | Attribute |
| 6 | > 12 Months | Binominal | Label |
| 7 | Job Title | Polynominal | Attribute |

### 3.3 Data Processing

Data that has been pre-proccessed can then be continued to be processed using the rapid miner application. Data processing was started by applying the K-Means clustering algorithm, then continued with the application of the Decision Tree algorithm and Naïve Bayes. After obtaining the results of the processing, we continue to analyze the results, so that the nature of this research is semiautomatic.

## 4. Results and Discussion

### 4.1 Implementation of the K-MEANS Algorithm

At this stage the data will be processed using the K-Means algorithm, the purpose of which is to group data according to the same characteristics to the same area and data with different characteristics to another region.

The first parameter we set is to determine the number of classes is 5, and will be tested 10 times to ensure consistent results. K-Means will automatically detect and analyze data and will try to classify existing data into several groups, where the data in one group have the same characteristics as each other and have different characteristics from the data in other groups.

### 4.2 Implementation of the Naïve Bayes Algorithm and Decision Tree

At this stage, the Decision Tree and Naïve Bayes algorithms are applied. Processing using two algorithms is intended to get two different processing results, which will then be compared the level of accuracy. 10 tests were conducted to get more accurate results (number of folds: 10) using random or automatic sample data.

The overall processing chart in this study is as follows:

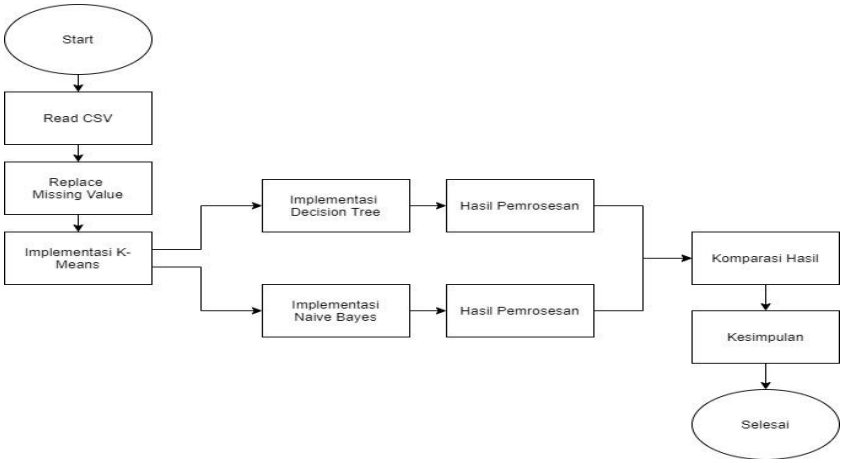


**Fig 3.** Implementation of the Naïve Bayes Algorithm and Decision Tree

### 4.3 Results of Data Mapping Based on K-Means Algorithm Processing

From the processing that has been done, the following results are obtained:

The result is that there are 5 clusters, with each mapping, according to data type and data type. With the following details:



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

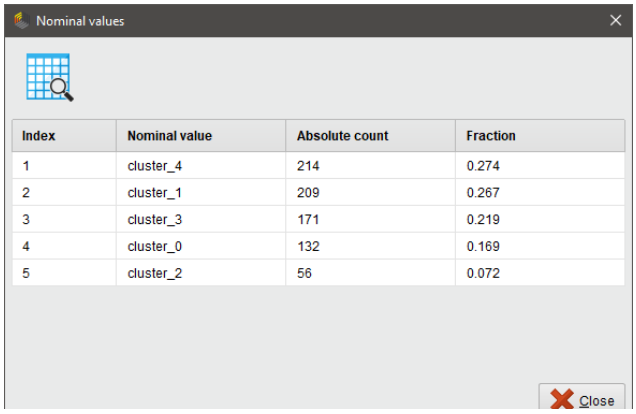


**Fig 4.** Results of Cluster Data Mapping

The following are the results of the data mapping. Mapping is based on gender, duration, and job title.

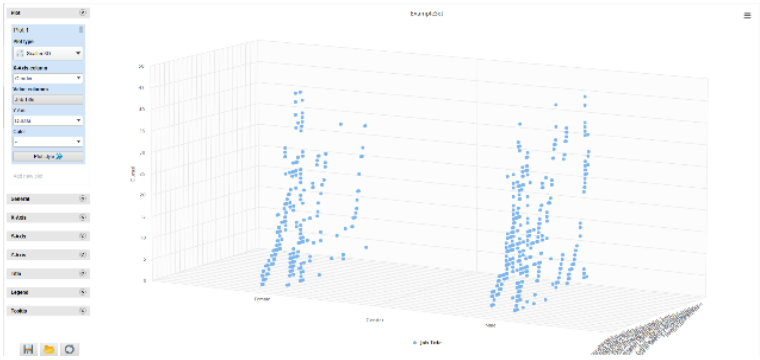

**Fig 5.** Results of Data Mapping Based on K-Means Algorithm Processing

### 4.4 Results of Data Analysis based on Decision Tree processing

After getting the data mapping from previous processing, the data analysis continues using the Decision Tree algorithm and Naïve Bayes. here are the results of the processing starting from the decision tree.

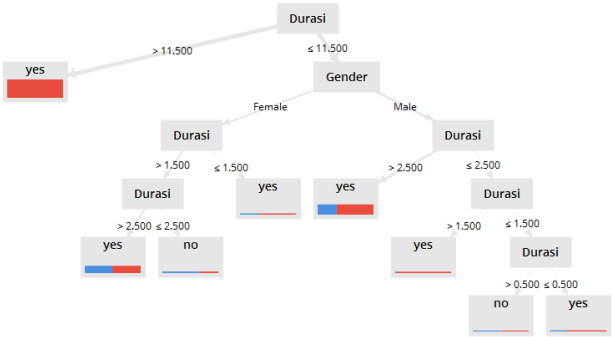


**Fig 6.** Results of Data Analysis



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

With the following information:

```
Durasi > 11.500: yes (no=0, yes=387)
Durasi ≤ 11.500
|   Gender = Female
|   |   Durasi > 1.500
|   |   |   Durasi > 2.500: yes (no=70, yes=72)
|   |   |   Durasi ≤ 2.500: no (no=10, yes=5)
|   |   Durasi ≤ 1.500: yes (no=3, yes=6)
|   Gender = Male
|   |   Durasi > 2.500: yes (no=71, yes=137)
|   |   Durasi ≤ 2.500
|   |   |   Durasi > 1.500: yes (no=0, yes=12)
|   |   |   Durasi ≤ 1.500
|   |   |   |   Durasi > 0.500: no (no=1, yes=1)
|   |   |   |   Durasi ≤ 0.500: yes (no=2, yes=5)
```

**Fig 7.** Results Information Data Analysis

### 4.5 Results of Data Analysis based on Naïve Bayes processing

The results of processing using the Naïve Bayes algorithm are shown in the following graphic diagram;

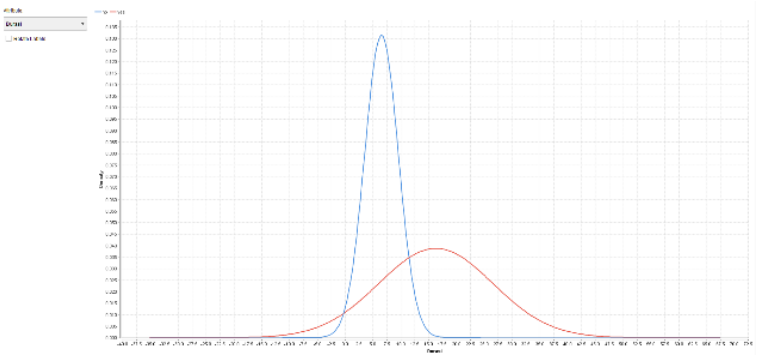

**Fig 8.** Results of Data Analysis based on Naive Bayes Processing

If we enlarge it for the intersection of the points, it will look like this

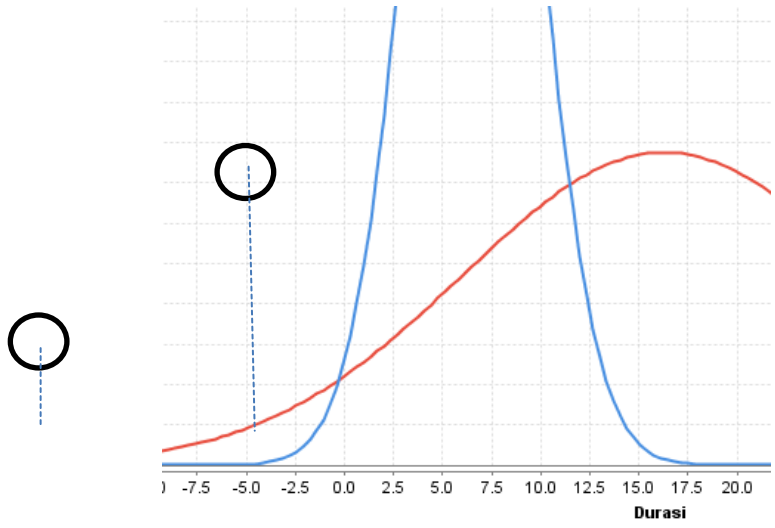


**Fig 9.** Data Analysis Results Intersection

At the intersection of the first point shows data with a duration of 0 months as the starting point for the two graphs, namely between cases> 12 is the same as yes or no. At the intersection of the second point it shows the starting point for cases> 12 months is the same as yes.

### 4.6 Comparison of Data Analysis Results based on Decision Tree and Naïve Bayes processing

From the results of the processing of the two algorithms previously described, the results will be compared between processing using the Decision Tree algorithm and Naïve Bayes. What will be the



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

comparison is the result of the Confusion Matrix. The Confusion Matrix presents the predictions and actual conditions of the data generated. In the Confusion Matrix there are true positive predictive values, false positive predictive values, true negative predictive values and false negative predictive values. In this comparison the Accuracy value will be calculated,*Precision*, and Recall.

a) *Accuracy* is a description of how accurately the model can classify correctly.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

b) *Precision* is the level of accuracy between the requested data and the prediction results provided by the model.

$$precision = \frac{TP}{TP + FP}$$

c) *Recall* is the ratio that predicts positive true values compared to overall positive true values.

$$recall = \frac{TP}{TP + FN}$$

### 4.7 Calculations on the Decision Tree algorithm

The following is the data processed by Rapid Miner.

```
accuracy: 91.69% +/- 2.00% (micro average: 91.69%)
ConfusionMatrix:
True:   no      yes
no:     138     46
yes:    19      579
```

**Fig 10.** Calculations on the Decision Algorithm

From the data above, it can be concluded that the performance vector is as follows

**Table 2.**
Conclusion Decision Tree Algorithm data

| | Yes | No. |
|---|---|---|
| Yes | 579 | 19 |
| No. | 46 | 138 |

*a) Accuracy*

$$\frac{579 + 138}{579 + 138 + 19 + 46} \times 100 = 91{,}687\ \%$$

*b) Precision*

$$\frac{579}{579 + 19} \times 100 = 96{,}82\ \%$$

*c) Recall class*

$$\frac{579}{579 + 46} \times 100 = 92{,}64\ \%$$

From these calculations it can be concluded that the level of accuracy produced by the Decision Tree algorithm processing is 91.687%

### 4.8 Calculations on the Naïve Bayes algorithm

The following is the data processed by Rapid Miner.

```
accuracy: 77.87% +/- 1.97% (micro average: 77.88%)
ConfusionMatrix:
True:   no      yes
no:     49      65
yes:    108     560
```

**Fig 11.** Rapid Miner Processing Data



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

From the data above, it can be concluded that the performance vector is as follows

**Table 3.**
Conclusion Naïve Bayes Algorithm

| | Yes | No. |
|---|---|---|
| Yes | 560 | 108 |
| No. | 65 | 49 |

*a) Accuracy*

$$\frac{560 + 49}{560 + 49 + 108 + 65} \times 100 = 77{,}877\ \%$$

*b) Precision*

$$\frac{560}{560 + 108} \times 100 = 83{,}832\ \%$$

*c) Recall class*

$$\frac{560}{560 + 65} \times 100 = 89{,}6\ \%$$

From these calculations it can be concluded that the level of accuracy produced by the Naïve Bayes algorithm processing is 77.877%

## 5. Conclusion

This article presents the results of an analysis of employee turnover using the Decision Tree and Naïve Bayes algorithms, where previously the data were clustered using the K-Means method. Clustering is determined by the number of clusters 5 and testing 10 times. Then from the results of the data processing, it can be seen that the accuracy produced by the Decision Tree algorithm is better at 91.69%, while the accuracy of Naïve Bayes is 77.87%.

For the turnover case, the results of this trial still contain weaknesses, one of which is the data attribute, where the available data has only a few attributes for processing, where in fact there are other factors that can affect turnover, such as education.

The author hopes that the results of this research can be used as a reference for readers, who will conduct similar research using broader and more complex data. So that this research continues and continues to grow.

## 6. Reference

[1] Pekerjaan, D. A. N. P., Kerja, K., Pada, K., Enseval, P. T., Manajemen, J., & Ekonomi, F. (2016). Analisis Pengaruh Turnover Karyawan, Motivasi, Dan Pengayaan Pekerjaan, Terhadap Kepuasan Kerja Karyawan Pada Pt. Enseval Megatrading Tbk Manado. *Jurnal Berkala Ilmiah Efisiensi*, *16*(3), 419–426.

[2] Rismayanti, R. D., Musadieq, M. Al, & Aini, E. K. (2018). Pengaruh Kepuasan Kerja Terhadap Turnover Intention Serta Dampaknya Pada Kinerja Karyawan. *Jurnal Administrasi Bisnis*, *61*(2), 127–136.

[3] Nurajijah, N., & Riana, D. (2019). Algoritma Naïve Bayes, Decision Tree, dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah. *Jurnal Teknologi Dan Sistem Komputer*, *7*(2), 77. https://doi.org/10.14710/jtsiskom.7.2.2019.77-82

[4] Gao, X., Wen, J., & Zhang, C. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering*, *2019*. https://doi.org/10.1155/2019/4140707

[5] Poerwanto, B., Palopo, U. C., & Yanu, R. (2016). *Analisis Cluster Menggunakan Algoritma K-Means*. (October).

[6] Chettri, R., Pradhan, S., & Chettri, L. (2015). Internet of Things: Comparative Study on Classification Algorithms (k-NN, Naive Bayes and Case based Reasoning). *International Journal of Computer Applications*, *130*(12), 7–9. https://doi.org/10.5120/ijca2015907120

[7] Yunita, D. (2017). Perbandingan Algoritma K-Nearest Neighbor dan Decision Tree untuk Penentuan Risiko Kredit Kepemilikan Mobil. *Jurnal Informatika Universitas Pamulang*, *2*(2), 103. https://doi.org/10.32493/informatika.v2i2.1512

[8] Ashari, A., Paryudi, I., & Min, A. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *International Journal of Advanced Computer Science and Applications*, *4*(11), 33–39. https://doi.org/10.14569/ijacsa.2013.041105

[9] Cardoni, F., & Cappella, E. (2018). *A COMPARISON OF DECISION TREE & NAIVE BAYES CLASSIFIERS AS SPAM FILTERING ALGORITHMS Optimization - DTC*. (January).



**Accredited "Rank 4" (Sinta 4),**DIKTI, No. 36 / E / KPT / 2019, December 13th 2019.

[10] Atma, U., Yogyakarta, J., Windarto, A. P., Tinggi, S., Komputer, I., Bangsa, T., & Wanto, A. (2019). *Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air*. (November). https://doi.org/10.30645/senaris.v1i0.81

[11] Yunmeng, Z., & Chengyi, Z. (2019). The Application of the Decision Tree Algorithm Based on K-means in Employee Turnover Prediction. *Journal of Physics: Conference Series*, *1325*(1). https://doi.org/10.1088/1742-6596/1325/1/012123

[12] Nengsih, W. (2017). Naive Bayes and decision tree modelling for comparative analysis method. *Journal of Engineering and Applied Sciences*, *12*(Specialissue4), 6672–6674. https://doi.org/10.3923/jeasci.2017.6672-6674

[13] Hairani, H., Nugraha, G. S., Abdillah, M. N., & Innuddin, M. (2018). Komparasi Akurasi Metode Correlated Naive Bayes Classifier dan Naive Bayes Classifier untuk Diagnosis Penyakit Diabetes. *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *3*(1), 6–11. https://doi.org/10.30743/infotekjar.v3i1.558

[14] Qowidho, T., Zarlis, M., Nababan, E. B., Agusnady, A., & Sembiring, B. S. (2019). Analysis of c4.5 and ID3 Methods in Determining Student Graduation. *Journal of Physics: Conference Series*, *1255*(1). https://doi.org/10.1088/1742-6596/1255/1/012025

[15] Parapat, J. S., & Sinaga, A. S. (2018). *Data Mining Algoritma C4 . 5 Pada Klasifikasi Kredit Koperasi Simpan Pinjam*. *4*(2).

[16] Rifai, A. I., Kerja, S., Jalan, P., Hambatan, B., Priok, T., Besar, B., … Algoritmi, C. (2015). Implementasi Data Mining Untuk Mendukung Sistem Manajemen Perkerasan Jalan Di Indonesia. Jurnal HPJI, 1(2), 93–104. https://doi.org/10.26593/.v1i2.1473.

[17] Olson, D. L., & Delen, D. (2008). Advanced data mining techniques. In Advanced Data Mining Techniques. https://doi.org/10.1007/978-3-540-76917-0

[18] Olson, D. L., & Delen, D. (2008). Advanced data mining techniques. In Advanced Data Mining Techniques. https://doi.org/10.1007/978-3-540-76917-0

[19] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan-Kaufmann Publishers, San Francisco, 2001.

[20] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 2012.